1 Corpus analysis and linguistic theory

When the first computer corpus, the Brown Corpus, was being created in the early 1960s, generative grammar dominated linguistics, and there was little tolerance for approaches to linguistic study that did not adhere to what generative grammarians deemed acceptable linguistic practice. As a consequence, even though the creators of the Brown Corpus, W. Nelson Francis and Henry Kučera, are now regarded as pioneers and visionaries in the corpus linguistics community, in the 1960s their efforts to create a machine-readable corpus of English were not warmly accepted by many members of the linguistic community. W. Nelson Francis (1992: 28) tells the story of a leading generative grammarian of the time characterizing the creation of the Brown Corpus as "a useless and foolhardy enterprise" because "the only legitimate source of grammatical knowledge" about a language was the intuitions of the native speaker, which could not be obtained from a corpus. Although some linguists still hold to this belief, linguists of all persuasions are now far more open to the idea of using linguistic corpora for both descriptive and theoretical studies of language. Moreover, the division and divisiveness that has characterized the relationship between the corpus linguist and the generative grammarian rests on a false assumption: that all corpus linguists are descriptivists, interested only in counting and categorizing constructions occurring in a corpus, and that all generative grammarians are theoreticians unconcerned with the data on which their theories are based. Many corpus linguists are actively engaged in issues of language theory, and many generative grammarians have shown an increasing concern for the data upon which their theories are based, even though data collection remains at best a marginal concern in modern generative theory.

To explain why corpus linguistics and generative grammar have had such an uneasy relationship, and to explore the role of corpus analysis in linguistic theory, this chapter first discusses the goals of generative grammar and the three types of adequacy (observational, descriptive, and explanatory) that Chomsky claims linguistic descriptions can meet. Investigating these three types of adequacy reveals the source of the conflict between the generative grammarian and the corpus linguist: while the generative grammarian strives for explanatory adequacy (the highest level of adequacy, according to Chomsky), the corpus linguist aims for descriptive adequacy (a lower level of adequacy), and it is arguable whether explanatory adequacy is even achievable through corpus analysis. However, even though generative grammarians and corpus linguists have different goals, it is wrong to assume that the analysis of corpora has nothing to contribute to linguistic theory: corpora can be invaluable resources for testing out linguistic hypotheses based on more functionally based theories of grammar, i.e. theories of language more interested in exploring language as a tool of communication. And the diversity of text types in modern corpora makes such investigations quite possible, a point illustrated in the middle section of the chapter, where a functional analysis of coordination ellipsis is presented that is based on various genres of the Brown Corpus and the International Corpus of English. Although corpora are ideal for functionally based analyses of language, they have other uses as well, and the final section of the chapter provides a general survey of the types of linguistic analyses that corpora can help the linguist conduct and the corpora available to carry out these analyses.

1.1 Linguistic theory and description

Chomsky has stated in a number of sources that there are three levels of "adequacy" upon which grammatical descriptions and linguistic theories can be evaluated: *observational* adequacy, *descriptive* adequacy, and *explanatory* adequacy.

If a theory or description achieves observational adequacy, it is able to describe which sentences in a language are grammatically well formed. Such a description would note that in English while a sentence such as *He studied for the exam* is grammatical, a sentence such as **studied for the exam* is not. To achieve descriptive adequacy (a higher level of adequacy), the description or theory must not only describe whether individual sentences are well formed but in addition specify the abstract grammatical properties making the sentences well formed. Applied to the previous sentences, a description at this level would note that sentences in English require an explicit subject. Hence, **studied for the exam* is ungrammatical and *He studied for the exam* is grammatical. The highest level of adequacy is explanatory adequacy, which is achieved when the description or theory not only reaches descriptive adequacy but does so using abstract principles which can be applied beyond the language being considered and become a part of "Universal Grammar." At this level of adequacy, one would describe the inability of English to omit subject pronouns as a consequence of the fact that, unlike Spanish or Japanese, English is not a language which permits "pro-drop," i.e. the omission of a subject pronoun that is recoverable from the context or deducible from inflections on the verb marking the case, gender, or number of the subject.

Within Chomsky's theory of principles and parameters, pro-drop is a consequence of the "null-subject parameter" (Haegeman 1991: 17–20). This parameter is one of many which make up universal grammar, and as speakers acquire a language, the manner in which they set the parameters of universal grammar is determined by the norms of the language they are acquiring. Speakers acquiring

English would set the null-subject parameter to negative, since English does not permit pro-drop; speakers of Italian, on the other hand, would set the parameter to positive, since Italian permits pro-drop (Haegeman 1991: 18).

Because generative grammar has placed so much emphasis on universal grammar, explanatory adequacy has always been a high priority in generative grammar, often at the expense of descriptive adequacy: there has never been much emphasis in generative grammar in ensuring that the data upon which analyses are based are representative of the language being discussed, and with the notion of the ideal speaker/hearer firmly entrenched in generative grammar, there has been little concern for variation in a language, which traditionally has been given no consideration in the construction of generative theories of language. This trend has become especially evident in the most recent theory of generative grammar: minimalist theory.

In minimalist theory, a distinction is made between those elements of a language that are part of the "core" and those that are part of the "periphery." The core is comprised of "pure instantiations of UG" and the periphery "marked exceptions" that are a consequence of "historical accident, dialect mixture, personal idiosyncracies, and the like" (Chomsky 1995: 19–20). Because "variation is limited to nonsubstantive elements of the lexicon and general properties of lexical items" (Chomsky 1995: 170), those elements belonging to the periphery of a language are not considered in minimalist theory; only those elements that are part of the core are deemed relevant for purposes of theory construction. This idealized view of language is taken because the goal of minimalist theory is "a theory of the initial state," that is, a theory of what humans know about language "in advance of experience" (Chomsky 1995: 4) before they encounter the real world of the language they are acquiring and the complexity of structure that it will undoubtedly exhibit.

This complexity of structure, however, is precisely what the corpus linguist is interested in studying. Unlike generative grammarians, corpus linguists see complexity and variation as inherent in language, and in their discussions of language, they place a very high priority on descriptive adequacy, not explanatory adequacy. Consequently, corpus linguists are very skeptical of the highly abstract and decontextualized discussions of language promoted by generative grammarians, largely because such discussions are too far removed from actual language usage. Chafe (1994: 21) sums up the disillusionment that corpus linguists have with purely formalist approaches to language study, noting that they "exclude observations rather than . . . embrace ever more of them" and that they rely too heavily on "notational devices designed to account for only those aspects of reality that fall within their purview, ignoring the remaining richness which also cries out for understanding." The corpus linguist embraces complexity; the generative grammarian pushes it aside, seeking an ever more restrictive view of language.

Because the generative grammarian and corpus linguist have such very different views of what constitutes an adequate linguistic description, it is clear why these two groups of linguists have had such a difficult time communicating and valuing each other's work. As Fillmore (1992: 35) jokes, when the corpus linguist asks the theoretician (or "armchair linguist") "Why should I think that what you tell me is *true*?", the generative grammarian replies back "Why should I think that what you tell me is *interesting*?" (emphasis added). Of primary concern to the corpus linguist is an accurate description of language; of importance to the generative grammarian is a theoretical discussion of language that advances our knowledge of universal grammar.

Even though the corpus linguist places a high priority on descriptive adequacy, it is a mistake to assume that the analysis of corpora has nothing to offer to generative theory in particular or to theorizing about language in general. The main argument against the use of corpora in generative grammar, Leech (1992) observes, is that the information they yield is biased more towards performance than competence and is overly descriptive rather than theoretical. However, Leech (1992: 108) argues that this characterization is overstated: the distinction between competence and performance is not as great as is often claimed, "since the latter is the product of the former." Consequently, what one discovers in a corpus can be used as the basis for whatever theoretical issue one is exploring. In addition, all of the criteria applied to scientific endeavors can be satisfied in a corpus study, since corpora are excellent sources for verifying the falsifiability, completeness, simplicity, strength, and objectivity of any linguistic hypothesis (Leech 1992: 112–13).

Despite Leech's claims, it is unlikely that corpora will ever be used very widely by generative grammarians, even though some generative discussions of language have been based on corpora and have demonstrated their potential for advancing generative theory. Working within the framework of government and binding theory (the theory of generative grammar preceding minimalist theory), Aarts (1992) used sections of the corpus housed at the Survey of English Usage at University College London to analyze "small clauses" in English, constructions like *her happy* in the sentence *I wanted her happy* that can be expanded into a clausal unit (*She is happy*). By using the London Corpus, Aarts (1992) was not only able to provide a complete description of small clauses in English but to resolve certain controversies regarding small clauses, such as establishing the fact that they are independent syntactic units rather than simply two phrases, the first functioning as direct object and the second as complement of the object.

Haegeman (1987) employed government and binding theory to analyze empty categories (i.e. positions in a clause where some element is missing) in a specific genre of English: recipe language. While Haegeman's investigation is not based on data from any currently available corpus, her analysis uses the type of data quite commonly found in corpora. Haegeman (1987) makes the very interesting claim that parametric variation (such as whether or not a language exhibits pro-drop) does not simply distinguish individual languages from one another but can be used to characterize regional, social, or register variation within a particular language. She looks specifically at examples from the genre (or register) of recipe language that contain missing objects (marked by the letters [a], [b], etc. in the example below):

(1) Skin and bone chicken, and cut [a] into thin slices. Place [b] in bowl with mushrooms. Purée remaining ingredients in blender, and pour [c] over chicken and mushrooms. Combine [d] and chill [e] well before serving. (Haegeman 1987: 236–7)

Government and binding theory, Haegeman (1987: 238) observes, recognizes four types of empty categories, and after analyzing a variety of different examples of recipe language, Haegeman concludes that this genre contains one type of empty category, *wh*-traces, not found in the core grammar of English (i.e. in other genres or regional and social varieties of English).

What distinguishes Haegeman's (1987) study from most other work in generative grammar is that she demonstrates that theoretical insights into universal grammar can be obtained by investigating the periphery of a language as well as the core. And since many corpora contain samples of various genres within a language, they are very well suited to the type of analysis that Haegeman (1987) has conducted. Unfortunately, given the emphasis in generative grammar on investigations of the core of a language (especially as reflected in Chomsky's recent work in minimalism), corpora will probably never have much of a role in generative grammar. For this reason, corpora are much better suited to functional analyses of language: analyses that are focused not simply on providing a formal description of language but on describing the use of language as a communicative tool.

1.2 Corpora in functional descriptions of language

Even though there are numerous functional theories of language, all have a similar objective: to demonstrate how speakers and writers use language to achieve various communicative goals.[1]

Because functionalists are interested in language as a communicative tool, they approach the study of language from a markedly different perspective than the generative grammarian. As "formalists," generative grammarians are primarily interested in describing the form of linguistic constructions and using these descriptions to make more general claims about Universal Grammar. For instance, in describing the relationship between *I made mistakes*, a sentence in the active voice, and its passive equivalent, *Mistakes were made by me*, a generative grammarian would be interested not just in the structural changes in word order between actives and passives in English but in making more general claims about the movement of constituents in natural language. Consequently, the movement of noun phrases in English actives and passives is part

[1] Newmeyer (1998: 13–18) provides an overview of the approaches to language study that various functional theories of language take.

of a more general process termed "NP [noun phrase] – movement" (Haegeman 1991: 270–3). A functionalist, on the other hand, would be more interested in the communicative potential of actives and passives in English. And to study this potential, the functionalist would investigate the linguistic and social contexts favoring or disfavoring the use of, say, a passive rather than an active construction. A politician embroiled in a scandal, for instance, might choose to utter the agentless passive *Mistakes were made* rather than *I made mistakes* or *Mistakes were made by me* because the agentless passive allows the politician to admit that something went wrong but at the same time to evade responsibility for the wrong-doing by being quite imprecise about exactly who made the mistakes.

Because corpora consist of texts (or parts of texts), they enable linguists to contextualize their analyses of language; consequently, corpora are very well suited to more functionally based discussions of language. To illustrate how corpora can facilitate functional discussions of language, this section contains an extended discussion of a functional analysis of elliptical coordinations in English based on sections of the Brown Corpus and the American component of the International Corpus of English (ICE). The goal of the analysis (described in detail in Meyer 1995) was not simply to describe the form of elliptical coordinations in speech and writing but to explain why certain types of elliptical coordinations are more common than others, why elliptical coordinations occur less frequently in speech than in writing, and why certain types of elliptical coordinations are favored more in some written genres than others.

The study was based on a 96,000-word corpus containing equal proportions of different types of speech and writing: spontaneous dialogues, legal cross examinations, press reportage, belles lettres, learned prose, government documents, and fiction. These genres were chosen because they are known to be linguistically quite different and to have differing functional needs. Government documents, for instance, are highly impersonal. Consequently, they are likely to contain linguistic constructions (such as agentless passives) that are associated with impersonality. Spontaneous dialogues, on the other hand, are much more personal, and will therefore contain linguistic constructions (such as the personal pronouns *I* and *we*) advancing an entirely opposite communicative goal. By studying genres with differing functional needs, one can take a particular linguistic construction (such as an elliptical coordination), determine whether it has varying frequencies and uses in different genres, and then use this information to determine why such distributions exist and to isolate the function (or communicative potential) of the construction.

In an elliptical coordination, some element is left out that is recoverable within the clause in which the ellipsis occurs. In the sentence *I wrote the introduction and John the conclusion* the verb *wrote* is ellipted in the second clause under identity with the same verb in the first clause. There are various ways to describe the different types of ellipsis occurring in English and other languages. Sanders (1977) uses alphabetic characters to identify the six different positions in which ellipsis can occur, ranging from the first position in the first clause (position A) to the last position in the second clause (position F):

A B C & D E F

Although there is disagreement about precisely which positions permit ellipsis in English, most would agree that English allows ellipsis in positions C, D, and E. Example (2) illustrates C-Ellipsis: ellipsis of a constituent at the end of the first clause (marked by brackets) that is identical to a constituent (placed in italics) at the end of the second clause.

(2) The author wrote [ ] and the copy-editor revised *the introduction to the book*.

Examples (3) and (4) illustrate D- and E-Ellipsis: ellipsis of, respectively, the first and second parts of the second clause.

(3) *The students* completed their course work and [ ] left for summer vacation.
(4) Sally *likes* fish, and her mother [ ] hamburgers.

The first step in studying the functional potential of elliptical coordinations in English was to obtain frequency counts of the three types of elliptical coordinations in the samples of the corpus and to explain the frequency distributions found. Of the three types of ellipsis in English, D-Ellipsis was the most frequent, accounting for 86 percent of the elliptical coordinations identified in the corpus. In contrast, both C- and E-Ellipsis were very rare, occurring in, respectively, only 2 percent and 5.5 percent of the elliptical coordinations.[2] These frequency distributions are identical to those found by Sanders (1977) in a survey he conducted of the frequency of ellipsis types in a variety of different languages. For instance, Sanders (1977) found that while all of the languages of the world allow D-Ellipsis, far fewer permit C-Ellipsis.

To explain typological distributions such as this, Sanders (1977) invokes two psycholinguistic constraints: the suspense effect (as Greenbaum and Meyer 1982 label it) and the serial position effect. Briefly, the suspense effect predicts that ellipsis will be relatively undesirable if the site of ellipsis precedes the antecedent of ellipsis, since the suspense created by the anticipation of the ellipted item places a processing burden on the hearer or reader. C-Ellipsis is therefore a relatively undesirable type of ellipsis because the antecedent of ellipsis (*the introduction to the book* in example 2) comes after the ellipsis in position C at the end of the first clause. D- and E-Ellipsis, on the other hand, are more desirable than C-Ellipsis because neither ellipsis type violates the suspense effect: for both types of ellipsis, the antecedent of ellipsis occurs in the first clause (position A for D-Ellipsis and position B for E-Ellipsis) in positions prior to ellipsis in the D- and E-positions in the second clause.

[2] The remaining 6.5 percent of elliptical coordinations consisted of constructions exhibiting more than one type of ellipsis and therefore no tendency towards any one type of ellipsis. For example, the example below contains both C- and D-Ellipsis: ellipsis of the direct object in the first clause and subject of the second clause.

(i) *We*$_1$ tried out [ ]$_2$ and [ ]$_1$ then decided to buy *the car*$_2$.

Table 1.1 *The favorability of C-, D-, and E- Ellipsis*

Ellipsis type	Suspense effect	Serial position effect
D-Ellipsis	F	F
E-Ellipsis	F	L
C-Ellipsis	L	L
F = favorable		
L = less favorable		

The serial position effect is based on research demonstrating that when given memory tests, subjects will remember items placed in certain positions in a series better than other positions. For instance, subjects will recall items placed first in a series more readily and accurately than items placed in the middle of a series. The results of serial learning experiments can be applied to the six positions in a coordinated construction (A–F) and make predictions about which antecedent positions will be most or least conducive to memory retention and thus favor or inhibit ellipsis. Position A, the antecedent position for D-Ellipsis (see example 3), is the position most favorable for memory retention. Consequently, D-Ellipsis will be the most desirable type of ellipsis according to the serial position effect. The next most favorable position for memory is position B, the antecedent position for E-Ellipsis, making this type of ellipsis slightly less desirable than D-Ellipsis. And increasingly less desirable for memory retention is the F-position, the antecedent position for C-Ellipsis, resulting in this type of ellipsis being the least desirable type of ellipsis in English.

Working together, the Suspense and Serial Position Effects make predictions about the desirability of ellipsis in English, predictions that match exactly the frequency distributions of elliptical coordinations found in the corpora. Table 1.1 lists the three types of ellipsis in English and the extent to which they favorably or unfavorably satisfy the suspense and serial position effects. D-Ellipsis quite favorably satisfies both the suspense and serial position effects, a fact offering an explanation of why D-Ellipsis was the most frequent type of ellipsis in the corpus. While E-Ellipsis satisfies the suspense effect, it less favorably satisfies the serial position effect, accounting for its less frequent occurrence in the corpus than D-Ellipsis. However, E-Ellipsis was more frequent than C-Ellipsis, a type of ellipsis that satisfies neither the suspense nor the serial position effect and was therefore the least frequent type of ellipsis in the corpus.

While the suspense and serial position effects make general predictions about the favorability or unfavorability of the three ellipsis types in English, they fail to explain the differing distributions of elliptical coordinations in speech and writing and in the various genres of the corpora. In speech, of the constructions in which ellipsis was possible, only 40 percent contained ellipsis, with the remaining 60 percent containing the full unreduced form. In writing, in contrast, ellipsis was much more common: 73 percent of the constructions in which ellipsis was possible contained ellipsis, with only 27 percent containing the full unreduced form. To explain these frequency differences, it is necessary to investigate why repetition (rather than ellipsis) is more necessary in speech than in writing.

The role of repetition in speech is discussed extensively by Tannen (1989: 47–53), who offers a number of reasons why a construction such as (5) below (taken from a sample of speech in the American component of ICE) is more likely to occur in speech than in writing.

(5) Yeah so *we got* that and *we got* knockers and *we got* bratwurst and *we got* <unintelligible>wurst or kranzwurst or something I don't know. (ICE-USA-S1A-016)

In (5), there are four repetitions of a subject and verb (*we got*) in the D-position that could have been ellipted rather than repeated. But in this construction, repetition serves a number of useful purposes quite unique to speech. First, as Tannen (1989: 48) observes, the repetition allows the speaker to continue the flow of the discourse "in a more efficient, less energy-draining way" by enabling him/her to continue speaking without worrying about editing what is being said and getting rid of redundancies, a task that would greatly slow down the pace of speech. At the same time, repetition is beneficial to the hearer "by providing semantically less dense discourse" (p. 49), that is, discourse containing an abundance of old rather than new information. Moreover, repetition can create parallel structures (as it does in example 5), and as many researchers have noted, parallelism is a very common device for enhancing the cohesiveness of a discourse.

In addition to having a different distribution in speech and writing, elliptical coordinations also had different distributions in the various genres of writing that were investigated. If the genres of fiction and government documents are compared, very different patterns of ellipsis can be found. In fiction, D-Ellipsis constituted 98 percent of the instances of ellipsis that were found. In government documents, on the other hand, D-Ellipsis made up only 74 percent of the instances of ellipsis, with the remaining 26 percent of examples almost evenly divided between C-Ellipsis and E-Ellipsis.

The high incidence of D-Ellipsis in fiction can be explained by the fact that fiction is largely narration, and narrative action, as Labov (1972: 376) has shown, is largely carried forth in coordinate sentences. These sentences will often have as subjects the names of characters involved in the narrative action, and as these names are repeated, they will become candidates for D-Ellipsis. For instance, in example (6) below (which was taken from a sample of fiction in the Brown Corpus), the second sentence (containing two coordinated clauses) begins with reference to a male character (*He*) at the start of the first clause, a reference that is repeated at the start of the second clause, leading to D-Ellipsis rather than repetition of the subject. Likewise, the last two sentences (which also consist of coordinated clauses) begin with references to another character (*Virginia* initially and then *She*), which are repeated and ellipted in the D-positions of subsequent clauses.

(6) The days seemed short, perhaps because his routine was, each day, almost the same. *He* rose late and [ ] went down in his bathrobe and slippers to have breakfast either alone or with Rachel. *Virginia* treated him with attention and [ ] tried to tempt his appetite with special food: biscuits, cookies, candies – the result of devoted hours in the tiled kitchen. *She* would hover over him and, looking like her brother, [ ] anxiously watch the progress of Scotty's fork or spoon. (K01 610–80)

Although the government documents in the corpus contained numerous examples of D-Ellipsis, they contained many more examples of C-Ellipsis than the samples of fiction did. One reason that C-Ellipsis occurred more frequently in government documents is that this type of construction has a function well suited to government documents. As Biber (1988) has noted, the genre in which government documents can be found, official documents, has a strong emphasis on information, "almost no concern for interpersonal or affective content" (p. 131), and a tendency towards "highly explicit, text-internal reference" (p. 142).

Instances of C-Ellipsis quite effectively help government documents achieve these communicative goals. First of all, because government documents are so focused on content or meaning, they are able to tolerate the stylistic awkwardness of constructions containing C-Ellipsis. In example (7) below (taken from a government document in the Brown Corpus), there is a very pronounced intonation pattern created by the C-Ellipsis, resulting in pauses at the site of ellipsis and just prior to the ellipted construction that give the sentence a rather abrupt and awkward intonation pattern.

(7) Each applicant is required to own [ ] or have sufficient interest in *the property to be explored*. (H01 1980–90)

This awkwardness is tolerated in government documents because of the overriding concern in this genre for accuracy and explicitness. An alternative way to word (7) would be to not ellipt the noun phrase in the C-position but instead to pronominalize it at the end of the second clause:

(8) Each applicant is required to own *the property to be explored* or have sufficient interest in *it*.

However, even though this wording results in no confusion in this example, in general when a third-person pronoun is introduced into a discourse, there is the potential that its reference will be ambiguous. If, in the case of (7), ellipsis is used instead of pronominalization, there is no chance of ambiguity, since the constraints for ellipsis in English dictate that there be only one source for the ellipsis in this sentence (the noun phrase *the property to be explored* in the second clause). Consequently, through ellipsis rather than pronominalization, the communicative goal of explicitness in government documents is achieved.

The discussion of coordination ellipsis in this section provides further evidence that corpus-based analyses can achieve "explanatory adequacy": the results of the study establish a direct relationship between the frequency of the various types of elliptical coordinations across the languages of the world